$$dist\_d(PC_{ab}, PC_{cd}) = \frac{1}{M}\left[\sum_{j=1}^{M} dist\_d(prob(conf(PC_{ab}))_j, prob(conf(PC_{cd}))_j\right],$$

Note: j refers to the components of the vector prob(conf(...)). Alternatively one can use the metric:

$$\text{dist\_d}(PC_{ab},PC_{cd}) = \sum_{i=1}^{M} p_i(1-p_i)$$

where the summation is over the attributes and p is the vector of probabilities for the discrete attribute if we merge conf(PC_{ab}) and conf(PC_{cd}). In all cases distance values near 0 mean high similarity.

Replace the paragraph beginning at page 18, line 1 and extending through page 18 line 7 with the following paragraph:

Now one can compute continuous and discrete distances between super-clusters:

$$\text{dist\_c}(V_i,V_j) = \frac{1}{|V_i||V_j|}\sum_{W_{ab}\in V_i}\sum_{W_{cd}\in V_j}|W_{ab}||W_{cd}|dist\_c(PC_{ab}, PC_{cd})$$

$$\text{dist\_d}(V_i,V_j) = \frac{1}{|V_i||V_j|}\sum_{W_{ab}\in V_i}\sum_{W_{cd}\in V_j}|W_{ab}||W_{cd}|dist\_d(PC_{ab}, PC_{cd}).$$

There are two stopping criteria for the merging: 1) Merging until we get a certain number of clusters - K -given by user. 2) Merging until the distances are larger then a threshold - maxMergeThreshold - supplied by a user.

Remarks

The changes to the text are made to make the notations with respect to distance metrics consistent with each other. An attachment to this amendment contains the two replacement paragraphs with additions underlined and deletions bracketed.

Respectfully Submitted,

Stephen J. Schultz
Reg No. 29,108

Attachment: text showing changes with underlining and bracketing

Changes shown with underlining and bracketing

Paragraph beginning at page 17, line 12.

Here, μ is the mean of the cluster with index ab.

Distance between two discrete cluster attributes is :

dist_d(conf(PC_{ab}),conf(PC_{cd})) = | prob(conf(PC_{ab})) – prob(conf(PC_{cd})) | . One then must take into account all the M attributes that are discrete:

$$dist\_d(PC_{ab}, PC_{cd}) = \frac{1}{M}\left[\sum_{j=1}^{M} dist\_d(prob(conf(PC_{ab}))_j, prob(conf(PC_{cd}))_j\right],$$

Note: j refers to the components of the vector prob(conf(...)). Alternatively one can use the metric:

$$\text{dist\_d([prob(conf(PC}_{ab}\text{)),prob(conf(PC}_{cd}\text{))] } \underline{\text{PC}_{ab}, \text{PC}_{cd}}) = \sum_{i=1}^{M} p_i(1-p_i)$$

where the summation is over the attributes and p is the vector of probabilities for the discrete attribute if we merge conf(PC_{ab}) and conf(PC_{cd}). In all cases distance values near 0 mean high similarity.

Paragraph beginning page 18, line 1 and extending through page 18, line 7.

Now one can compute continuous and discrete distances between super-clusters:

$$\text{dist\_c}(V_i, V_j) = \frac{1}{|V_i||V_j|}\sum_{W_{ab}\in V_i}\sum_{W_{cd}\in V_j}|W_{ab}||W_{cd}|dist\_c(PC_{ab}, PC_{cd})$$

$$\text{dist\_d}(V_i, V_j) = \frac{1}{|V_i||V_j|}\sum_{W_{ab}\in V_i}\sum_{W_{cd}\in V_j}|W_{ab}||W_{cd}|dist\_d([conf(PC_{ab}), conf(PC_{cd}))]\ \underline{\text{PC}_{ab}, \text{PC}_{cd}}).$$

There are two stopping criteria for the merging: 1) Merging until we get a certain number of clusters - K -given by user. 2) Merging until the distances are larger then a threshold - maxMergeThreshold - supplied by a user.